Decentralized Q-Learning for Stochastic Teams and Games

Gürdal Arslan and Serdar Yüksel, *Member, IEEE*

***Abstract*—There are only a few learning algorithms applicable to stochastic dynamic teams and games which generalize Markov decision processes to decentralized stochastic control problems involving possibly self-interested decision makers. Learning in games is generally difficult because of the non-stationary environment in which each decision maker aims to learn its optimal decisions with minimal information in the presence of the other decision makers who are also learning. In stochastic dynamic games, learning is more challenging because, while learning, the decision makers alter the state of the system and hence the future cost. In this paper, we present decentralized Q-learning algorithms for stochastic games, and study their convergence for the weakly acyclic case which includes team problems as an important special case. The algorithm is decentralized in that each decision maker has access to only its local information, the state information, and the local cost realizations; furthermore, it is completely oblivious to the presence of other decision makers. We show that these algorithms converge to equilibrium policies almost surely in large classes of stochastic games.**

I. Introduction

This paper aims at developing new learning algorithms with desirable convergence properties for certain classes of stochastic games, which are discrete-time dynamic games in which the history can be summarized by a "state" [1]. More specifically, we focus on weakly acyclic stochastic games that can be used to model cooperative systems. The chief merit of the paper lies in the fact that learning takes place in stochastic games, which are truly dynamic games, as opposed to learning in repeated games in which the same single-stage game is played in every stage. In stochastic games, the policies selected by the decision makers not only impact their immediate cost but also alter the stage-games to be played in the future through the state dynamics. Hence, our results are applicable to a significantly broader set of applications.

The existing literature on learning in stochastic games is very small in comparison with the literature on learning in repeated games. As the method of reinforcement learning gained popularity in the context of Markov decision problems, a surge of interest in generalizing the method of reinforcement learning, in particular Q-learning algorithm [2], to stochastic games has led to a set of publications primarily in the computer science literature; see [3] and the references therein. In many of these publications, the authors tend to assume that the real objective of the agents [1] is for some reason to find and play an equilibrium strategy (and sometimes this even requires agents to somehow agree on a particular equilibrium strategy), and not necessarily to pursue their own objectives. Another serious issue is that the multi-agent algorithms introduced in many of these recent papers are not scalable since each agent needs to maintain estimates of its Q-factors for each state/joint action pair and compute an equilibrium at each step of the algorithm using the updated estimates, assuming that the actions and objectives are exchanged between all agents.

Standard Q-learning, which enables an agent to learn how to play optimally in a single-agent environment, has also been applied to very specific multi agent applications [4], [5]. Here, each agent runs a standard Q-learning algorithm by ignoring the other agents, and hence information exchange between agents and computational burden on each agent are substantially lower than aforementioned multi-agent extensions of Q-learning algorithm. Also, standard Q-learning in a multi-agent environment makes sense from individual bounded rationality point of view. However, no analytical results exist regarding the properties of standard Q-learning in a stochastic game setting.

We should also mention several attempts to extend a well-known learning algorithm called Fictitious Play (FP) [6], [7] to stochastic games [8], [9], [10]. The joint action learning algorithm presented in [8] would be computationally prohibitive quickly as the number of agents/states/actions grow. The algorithms presented in [8] are claimed to be convergent to an equilibrium in single-state single-stage common interest games but without a proof. The extension of FP considered in [9] requires each agent to calculate a stationary policy at each step in response to the empirical frequencies of the stationary policies calculated and announced by other agents in the past. The main contribution of [9] is to show that such FP algorithm is not convergent even in the simplest 2x2x2 stochastic game where there are two states and two agents with two moves for each agent. The version of FP used in [10] is applicable only to zero-sum games (strictly adversarial games).

Other related work includes [11], [12], [13]. In [11], a multi-agent version of an actor-critic algorithm [14] is shown to be convergent to generalized equilibria in a weak sense of convergence, whereas in [12] a policy iteration algorithm is presented without rigorous results for stochastic games. The algorithms given in [11], [12] are rational from individual agent perspective, however they require higher level of data storing and processing than standard Q-learning. The paper

Part of this paper is presented at the IEEE Conference on Decision and Control, December 2015. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

G. Arslan is with the Department of Electrical Engineering, University of Hawaii at Manoa, 440 Holmes Hall, 2540 Dole Street, Honolulu, HI 96822, USA. gurdal@hawaii.edu

S. Yüksel is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, CANADA, K7L 3N6. yuksel@mast.queensu.ca

[1] The terms "agent" and "decision maker" are used interchangeably.

[13] uses the policy iteration algorithm given in [12] in conjunction with certain approximation methods to deal with a large state-space in a specific card-game without rigorous results.

We should emphasize that our viewpoint is individual bounded rationality and strategic decision making, that is, agents should act to pursue their own objectives even in the short term using localized information and reasonable algorithms. It is also desired that agent strategies converge to an agreeable solution in cooperative situations where agent objectives are aligned with system designer's objective even though agents do not necessarily strive for converging to a particular strategy.

The rest of the paper is organized as follows. In §II, the model is introduced. In § III, the specifics of the learning paradigm and the standard Q-learning algorithm is discussed, followed by the presentation of our first Q-learning algorithm for stochastic games and its convergence properties. Generalizations of our main results in § III are presented in §IV. This is followed by a simulation study in §V. The paper is concluded with some final remarks in §VI. Appendices contain the proofs of the technical results in the paper.

II. Stochastic Dynamic Games

Consider the (discrete-time) networked control system illustrated in Figure 1 where x_t is the state of the system at time t, u_t^i is the input generated by controller i at time t, and w_t is the random disturbance input at time t. Suppose that each controller i is an autonomous decision maker (DM) interested in minimizing its own long-term cost

$$E\left[\sum_{t\geq 0} c^i(x_t, u_t^1, \ldots, u_t^N)\right]$$

where $c^i(x_t, u_t^1, \ldots, u_t^N)$ is the cost incurred by controller i at time t, and $E[\cdot]$ denotes the expectation given a collection of control policies (which will be specified later in the paper) on a probability space $(\Omega, \mathcal{F}, P)$. Although controller i can only choose its own decisions $u_0^i, u_1^i, \ldots$, its cost generally depends on the decisions of all controllers through its single-stage cost as well as the state dynamics. This dynamic coupling between self-interested DMs with long-term objectives naturally lead to the framework of stochastic games [1] which generalize Markov decision problems.

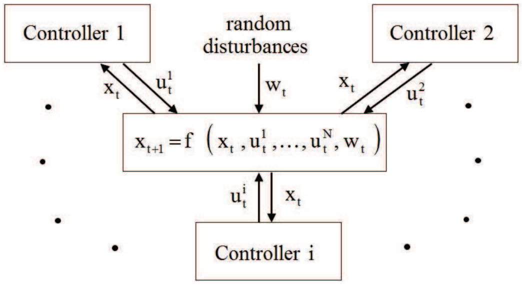

Fig. 1. A networked control system.

Over the past half-century, there have been many applications of stochastic games on control problems; see Chapter XIV in [15] as an early reference. At the present time, the control theory literature includes a large number of papers employing the theory of stochastic games and their continuous-time counterparts called "differential games" [16]. Many papers in this body of work study a zero-sum game between a controller which aims to optimize the system performance and an adversary which controls certain system parameters and inputs to make the system performance as poor as possible. We selectively cite [17] for robust control and minimax estimation problems, [18] for flow control in queueing networks, [19] for control of hybrid systems, and [20] for robustness, security, and resilience of cyber-physical control systems. The case of nonzero-sum games in which the decision makers do not always have diametrically opposed objectives has also received significant attention; see for example [21] on admission, service, and routing control in queueing systems, [22] on transmission control in cognitive radio systems, [23] on network security, and [24] on formation control.

We should also mention the work on team decision problems where all DMs share a common long-term objective albeit with access to different information variables; see e.g., [25], [26]. In this paper, differently from the usual team decision problems in the literature, even though each DM has access to the state information, it does not have access to global information on the other DMs, and even their presence. We also note that the emergence of distributed control systems requires the formulation of "team problems" within a game-theoretic framework where local controllers are tasked to achieve one system level objective without centralized coordination; see for example [27] on distributed model predictive control. This type of team problems and its generalizations where the objectives of DMs are aligned in some sense with a team objective are the primary focus of our work though the class of games considered in this paper is more general and it even includes some zero-sum stochastic games.

A. Discounted Stochastic Dynamic Games

A (finite) discounted stochastic game has the following ingredients; see [1].

- A finite set of DMs with the $i-$th DM referred to as DM^i for $i \in \{1, \ldots, N\}$
- a finite set $\mathbb{X}$ of states
- a finite set $\mathbb{U}^i$ of control decisions for each DM^i
- a cost function c^i for each DM^i determining DM^i's cost $c^i(x, u^1, \ldots, u^N)$ at each state $x \in \mathbb{X}$ and for each joint decision $(u^1, \ldots, u^N) \ \in \mathbb{U}^1 \times \cdots \mathbb{U}^N$
- a discount factor $\beta^i \in (0,1)$ for each DM^i
- a random initial state $x_0 \in \mathbb{X}$
- a transition kernel for the probability $P[x'|x, u^1, \ldots, u^N]$ of each state transition from $x \in \mathbb{X}$ to $x' \in \mathbb{X}$ for each joint decision $(u^1, \ldots, u^N) \in \mathbb{U}^1 \times \cdots \mathbb{U}^N$.

Such a stochastic game induces a discrete-time controlled Markov process where the state at time t is denoted by $x_t \in \mathbb{X}$ starting with the initial state x_0. At any time $t \geq 0$, each DM^i makes a control decision $u_t^i \in \mathbb{U}^i$ (possibly randomly) based on the available information. The state x_t and the joint decisions $(u_t^1,...,u_t^N)$ together determine each DM^i's cost $c^i(x_t, u_t^1,...,u_t^N)$ at time t as well as the probability distribution $P[\ \cdot \mid x_t, u_t^1,...,u_t^N]$ with which the next state x_{t+1} is selected.

A policy for a DM is a rule of choosing an appropriate control decision at any time based on the DM's history of observations. We will focus on stationary policies of the form where a DM's decision at time t is determined solely based on the state x_t. Such policies for each DM^i are identified by mappings from the state space $\mathbb{X}$ to the set $\mathcal{P}(\mathbb{U}^i)$ of probability distributions on $\mathbb{U}^i$. The interpretation is that a DM^i using such a policy $\pi^i : \mathbb{X} \mapsto \mathcal{P}(\mathbb{U}^i)$ makes its decision u_t^i at any time t by choosing randomly from $\mathbb{U}^i$ according to $\pi^i(x_t)$. We will denote the set of such policies by Δ^i for each DM^i. We will primarily be interested in deterministic (stationary) policies[2] denoted by Π^i for each DM^i, where each policy $\pi^i \in \Pi^i$ is identified by a mapping from $\mathbb{X}$ to $\mathbb{U}^i$.

The objective of each DM^i is to find a policy $\pi^i \in \Delta^i$ that minimizes its expected discounted cost

$$J_x^i(\pi^1, \dots \pi^N) = E_x\left[\sum_{t\geq 0}(\beta^i)^t c^i\left(x_t, u_t^1, \dots, u_t^N\right)\right] \qquad (1)$$

for all $x \in \mathbb{X}$, where E_x denotes the conditional expectation given $x_0 = x$. Since DMs have possibly different cost functions and each DM's cost may depend on the control decisions of the other DMs, we adopt the notion of equilibrium to represent those policies that are *person-by-person optimal*. For ease of notation, we denote the policies of all DMs other than DM^i by π^{-i}. For future reference, we also define $\Pi^{-i} := \times_{j\neq i}\Pi^j$ and $\Delta^{-i} := \times_{j\neq i}\Delta^j$ as well as $\Pi := \times_j \Pi^j$ and $\Delta := \times_j \Delta^j$. Using this notation, we write a joint policy $(\pi^1, \dots \pi^N)$ as (π^i, π^{-i}) and $J_x^i(\pi^1, \dots \pi^N)$ as $J_x^i(\pi^i, \pi^{-i})$.

Definition 1: A joint policy $(\pi^{*1}, \dots, \pi^{*N}) \in \Delta$ constitutes an (Markov perfect) equilibrium if, for all i, x,

$$J_x^i(\pi^{*i}, \pi^{*-i}) = \min_{\pi^i \in \Delta^i} J_x^i(\pi^i, \pi^{*-i}).$$

It is known that any finite discounted stochastic game possesses an equilibrium policy as defined above [28].

Although the minimum above can always be achieved by a deterministic policy in Π^i (since each DM^i's problem is a stationary Markov decision problem when the policies of the other DMs are fixed at π^{*-i}), a deterministic equilibrium policy may not exist in general. However, many interesting classes of games do possess equilibrium in deterministic policies. In particular, large classes of games arising from applications where all DMs benefit from cooperation possess equilibrium in deterministic policies. The primary examples of such games of cooperation are team problems where all DMs have the same cost function. In team problems, the deterministic policies minimizing the common cost function are clearly equilibrium policies although non-optimal deterministic equilibrium policies may also exist. A more general set of games of cooperation are those in which some function, called the potential function, decreases whenever a single DM decreases its own cost by unilaterally switching from one deterministic policy to another one. In this class of games, the deterministic policies minimizing the potential function are equilibrium policies. As such, we are primarily interested in the set of deterministic equilibrium policies denoted by Π_{eq}, where $\Pi_{\text{eq}} \subset \Pi$.

We next formally introduce the set of games considered in this paper.

B. Weakly Acyclic Games

Let $\Pi^i_{\pi^{-i}}$ denote DM^i's set of (deterministic) best replies to any $\pi^{-i} \in \Delta^{-i}$, i.e.,

$$\Pi^i_{\pi^{-i}} := \big\{\hat{\pi}^i \in \Pi^i : J_x(\hat{\pi}^i, \pi^{-i}) = \min_{\pi^i \in \Delta^i} J_x(\pi^i, \pi^{-i}), \\ \text{for all } x\big\}.$$

DM^i's best replies to any $\pi^{-i} \in \Delta^{-i}$ can be characterized by its optimal Q-factors $Q^i_{\pi^{-i}}$ satisfying the fixed point equation

$$\begin{aligned} Q^i_{\pi^{-i}}(x, u^i) =& E_{\pi^{-i}(x)}\big[c^i(x, u^i, u^{-i}) \\ &+ \beta^i \sum_{x' \in \mathbb{X}} P[x'|x, u^i, u^{-i}] \min_{v^i \in \mathbb{U}^i} Q^i_{\pi^{-i}}(x', v^i)\big] \end{aligned} \qquad (2)$$

for all x, u^i, where $E_{\pi^{-i}(x)}$ denotes the expectation with respect to the joint distribution of u^{-i} given by $\pi^{-i}(x) = \pi^1(x) \times \cdots \times \pi^{i-1}(x) \times \pi^{i+1}(x) \times \cdots \times \pi^N(x)$. The optimal Q-factor $Q^i_{\pi^{-i}}(x, u^i)$ represents DM^i's expected discounted cost to go from the initial state x assuming that DM^i initially chooses u^i and uses an optimal policy thereafter while the other DMs use π^{-i}. One can then write $\Pi^i_{\pi^{-i}}$ as

$$\Pi^i_{\pi^{-i}} = \big\{\hat{\pi}^i \in \Pi^i : Q^i_{\pi^{-i}}(x, \hat{\pi}^i(x)) = \min_{v^i \in \mathbb{U}^i} Q^i_{\pi^{-i}}(x, v^i), \\ \text{for all } x\big\}.$$

The set of (deterministic) joint best replies is denoted by $\Pi_\pi := \Pi^1_{\pi^{-1}} \times \cdots \times \Pi^N_{\pi^{-N}}$. Any best reply $\hat{\pi}^i \in \Pi^i_{\pi^{-i}}$ of DM^i is called a *strict best reply* with respect to (π^i, π^{-i}) if

$$J_x^i(\hat{\pi}^i, \pi^{-i}) < J_x^i(\pi^i, \pi^{-i}), \quad \text{for some } x.$$

Such a strict best reply $\hat{\pi}^i$ achieves DM^i's minimum cost given π^{-i} for all initial states and results in a strict improvement over π^i for at least one initial state.

Definition 2: We call a (possibly finite) sequence of deterministic joint policies $\pi_0, \pi_1, \dots$ a *strict best reply path* if, for each k, π_k and π_{k+1} differ in exactly one DM position, say DM^i, and π^i_{k+1} is a strict best reply with respect to π_k.

Definition 3: A discounted stochastic game is called *weakly acyclic* under strict best replies if there is a strict best reply path starting from each deterministic joint policy and ending at a deterministic equilibrium policy.

Figure 2 shows the strict best reply graph of a game where the nodes represent the deterministic joint policies and the directed edges represent the single-DM strict best replies. Each deterministic equilibrium policy is represented by a sink, i.e., a node with no outgoing edges, in such a graph. Note that the game illustrated in Figure 2 is weakly acyclic under strict best

[2] When it is not clear from the context, a “policy” will mean a deterministic policy.

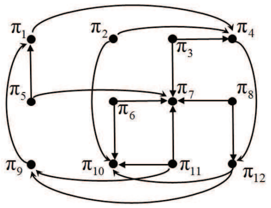

Fig. 2. The strict best reply graph of a stochastic game.

replies since there is a path from every node to a sink (π_7 or π_{10}). Note also that a weakly acyclic game may have cycles in its strict best reply graph, for example, $\pi_1 \to \pi_4 \to \pi_{12} \to \pi_9$ in Figure 2.

Weakly acyclic games constitute a fairly large class of games. In the case of single-stage games, all potential games as well as dominance solvable games are examples of weakly acyclic games; see [29]. We note that the concept of weak acyclicity introduced in this paper is with respect to the stationary Markov policies for stochastic games, and constitutes a generalization of weak acyclicity introduced in [30] for single-stage games. The primary examples of weakly acyclic games in the case of stochastic games are the team problems with finite state and control sets where DMs have identical cost functions and discount factors. Clearly, many other classes of stochastic games are weakly acyclic, e.g., appropriate multi-stage generalizations of potential games and dominance solvable games restricted to the stationary Markov policies are weakly acyclic for the same reason that the single-stage versions of these games are weakly acyclic [29].

C. A Best Reply Process for Weakly Acyclic Games

Consider a policy adjustment process in which only one DM updates its policy at each step by switching to one of its strict best replies. Such a process would terminate at an equilibrium policy if the game has no cycles in its strict best reply graph and the process continues until no DM has strict best replies. A weakly acyclic game may contain cycles in its strict best reply graph but there must be some edges leaving each cycle because otherwise there would not be a path from each node to a sink. Therefore, as long as each updating DM considers each of its strict best replies with positive probability, the adjustment process would terminate at an equilibrium policy in a weakly acyclic game with probability (w.p.) one. This adjustment process would require a criterion to determine the updating DM at each step and the DMs would have to a priori agree to this criterion. An equilibrium policy can be reached through a similar adjustment process without a pre-game agreement on the selection of the updating DM, if all DMs update their policies at each step but with some inertia. Consider now the following policy adjustment process, which is the best reply process with memory length of one and inertia introduced in Sections 6.4-6.5 of [30].

Best Reply Process with Inertia (for DM^i):

```
Set parameters
    $\lambda^i \in (0,1)$: inertia
Initialize $\pi_0^i \in \Pi^i$ (arbitrary)
Iterate $k \geq 0$
    If $\pi_k^i \in \Pi^i_{\pi_k^{-i}}$
        $\pi_{k+1}^i = \pi_k^i$
    Else
        $\pi_{k+1}^i = \begin{cases} \pi_k^i & \text{w.p. } \lambda^i \\ \text{any } \pi^i \in \Pi^i_{\pi_k^{-i}} & \text{w.p. } (1-\lambda^i)/\left|\Pi^i_{\pi_k^{-i}}\right| \end{cases}$
    End
```

On the one hand, if the joint policy $\pi_k := (\pi_k^1, \ldots, \pi_k^N)$ is an equilibrium policy at any step k, then the policies will never change in the subsequent steps. On the other hand, regardless of what the joint policy $\pi_k := (\pi_k^1, \ldots, \pi_k^N)$ is at any step k, the joint policy π_{k+L} in L steps later will be an equilibrium policy with positive probability $p_{\min} > 0$ where L is the maximum length of the shortest strict best reply path from any policy to an equilibrium policy and $p_{\min}$ depends only on the inertias $\lambda^1, \ldots, \lambda^N$, and L. This readily implies that the best reply process with inertia will reach an equilibrium policy in finite number of steps w.p. 1 [30], i.e.,

$$P[\pi_k = \pi^*, \text{ for some } \pi^* \in \Pi_{\text{eq}} \text{ and all large } k < \infty] = 1.$$

We now note that each updating DM^i at step k needs to compute its best replies $\Pi^i_{\pi_k^{-i}}$, which can be done by first solving the fixed point equation (2) for $\pi^{-i} = \pi_k^{-i}$. DM^i can solve (2), for example through value iterations, provided that DM^i knows the state transition probabilities P and the policies π_k^{-i} of the other DMs to evaluate the expectations in (2). In most realistic situations, DMs would not have access to such information and therefore would not be able to compute their best replies directly. In the next section, we introduce our learning paradigm in which DMs would be able to learn their near best replies with minimal information and adjust their policies (approximately) along the strict best reply paths as in the best reply process with inertia.

III. Q-Learning in Stochastic Dynamic Games

A. Learning Paradigm for Stochastic Dynamic Games

The learning setup involves specifying the information that DMs have access to. We assume that each DM^i knows its own set $\mathbb{U}^i$ of decisions and its own discount factor β^i. In addition, before choosing its decision u_t^i at any time t, each DM^i has the knowledge of

- its own past decisions $u_0^i, \ldots, u_{t-1}^i$, and
- past and current state realizations $x_0, \ldots, x_t$, and
- its own past cost realizations

$$c^i(x_0, u_0^i, u_0^{-i}), \ldots, c^i(x_{t-1}, u_{t-1}^i, u_{t-1}^{-i}).$$

Each DM^i has access to no other information such as the state transition probabilities or any information regarding the other DMs (not even the existence of the other DMs). In effect, the problem of decision making from the perspective of each DM^i appears to be a stationary Markov decision problem. It is reasonable that each DM^i with this view of its environment would use the standard Q-learning algorithm [2] to learn its optimal Q-factors and its optimal decisions. This would lead to the following Q-learning dynamics for each DM^i:

$$\begin{aligned} Q_{t+1}^i(x,u^i) =& Q_t^i(x,u^i), \quad \text{for all } (x,u^i) \neq (x_t,u_t^i) \\ Q_{t+1}^i(x_t,u_t^i) =& Q_t^i(x_t,u_t^i) + \alpha_t^i \big[c^i(x_t,u_t^i,u_t^{-i}) \\ &+ \beta^i \min_{v^i \in \mathbb{U}^i} Q_t^i(x_{t+1},v^i) - Q_t^i(x_t,u_t^i)\big] \end{aligned}$$

where $\alpha_t^i \in [0,1]$ denotes DM^i's step size at time t.

If only one DM, say DM^i, were to use Q-learning and the other DMs used constant policies π^{-i}, then DM^i would asymptotically learn its corresponding optimal Q-factors, i.e.,

$$P[Q_t^i \to Q_{\pi^{-i}}^i] = 1$$

provided that all state-control pairs x, u^i are visited infinitely often and the step sizes are reduced at a proper rate. This follows from the well-known convergence of Q-learning in a stationary environment; see [31]. To exploit the learnt Q-factors while maintaining exploration, the actual decisions are often selected with very high probability as

$$u_t^i \in \text{argmin}_{v^i \in \mathbb{U}^i} Q_t^i(x_t,v^i)$$

and with some small probability any decision in $\mathbb{U}^i$ is experimented. One common way of achieving this for DM^i is to select any decision $u^i \in \mathbb{U}^i$ randomly according to (Boltzman action selection)

$$P[u_t^i = u^i|\mathcal{F}_t] = \frac{e^{-Q_t^i(x_t,u^i)/\tau}}{\sum_{v^i \in \mathbb{U}^i} e^{-Q_t^i(x_t,v^i)/\tau}}$$

where $\tau > 0$ is a small constant called the temperature parameter, and $\mathcal{F}_t$ is the history of the random events realized up to the point just before the selection of $(u_t^1,\ldots,u_t^N)$.

However, when all DMs use Q-learning and select their decisions as described above, the environment is non-stationary for all DMs, and there is no reason to expect convergence in that case. In fact, one can construct examples where DMs using Q-learning are caught up in persistent oscillations; see Section 4 in [32] for the non-convergence of Q-learning in Shapley's game. However, the convergence of Q-learning may still be possible in team problems, coordination-type games, or more generally in weakly-acyclic games. It is instructive to first consider the repeated games.

Here, there is no state dynamics (the set $\mathbb{X}$ of states is a singleton) and the DMs have no look-ahead ($\beta^1 = \cdots \beta^N = 0$). The only dynamics in this case is due to Q-learning which reduces to the averaging dynamics

$$Q_{t+1}^i(u^i) = Q_t^i(u^i), \quad \text{for all } u^i \neq u_t^i \tag{3}$$

$$Q_{t+1}^i(u_t^i) = Q_t^i(u_t^i) + \alpha_t^i \left[c^i(u_t^i,u_t^{-i}) - Q_t^i(u_t^i)\right] \tag{4}$$

where

$$P[u_t^i = u^i|\mathcal{F}_t] = \frac{e^{-Q_t^i(u^i)/\tau}}{\sum_{v^i \in \mathbb{U}^i} e^{-Q_t^i(v^i)/\tau}}. \tag{5}$$

The long-term behavior of these averaging dynamics is analyzed in [32] and strongly connected to the long-term behavior of the well-known Stochastic Fictitious Play (SFP) dynamics [33] in the case of two DMs; see Lemma 4.1 in [32]. In two-DM SFP, each DM^i tracks the empirical frequencies of the past decisions of its opponent DM^{-i} and chooses a nearly optimal response (with some experimentation) based on the incorrect assumption that DM^{-i} will choose its decisions according to the empirical frequencies of its past decisions

$$q_t^{-i}(u^{-i}) = \frac{1}{t}\sum_{k=0}^{t-1} I_{\{u_t^{-i}=u^{-i}\}}, \quad \text{for all } u^{-i}$$

where $I_{\{\cdot\}}$ is the indicator function and

$$\begin{aligned} P[u_t^i = u^i|\mathcal{F}_t] =& \frac{e^{-M_t^i(u^i)/\tau}}{\sum_{v^i \in \mathbb{U}^i} e^{-M_t^i(v^i)/\tau}} \\ M_t^i(u^i) :=& \sum_{u^{-i}} q_t^{-i}(u^{-i}) c^i(u^i,u^{-i}). \end{aligned}$$

Using the connection between Q-learning dynamics (4)-(5) and SFP dynamics, the convergence of Q-learning (4)-(5) is established in zero-sum games as well as in partnership games with two DMs; see Proposition 4.2 in [32]. It may be possible to extend this convergence result to multi-DM potential games [34], [35], but this is currently unresolved. However, given the nonconvergence of FP (where DMs choose exact optimal responses with no experimentation, i.e., $\tau \downarrow 0$) in some coordination games [36], the prospect of establishing the convergence of Q-learning even in all two-DM weakly acyclic games does not seem promising.

It is possible to employ additional features such as the truncation of the observation history or multi-time-scale learning to obtain learning dynamics that are convergent in all repeated weakly acyclic games; see our own previous work [37] and the others [38], [30], [39], [40]. However, the question of learning an equilibrium policy in stochastic games is an open question. The only relevant reference considering the stochastic games is [11] where each DM uses value learning coupled with policy search at a slower time-scale. The results in [11] apply to all stochastic games and therefore they are necessarily quite weak. Loosely speaking, the main result in [11] shows that the limit points of certain empirical measures (weighted with the step sizes) in the policy space constitute "generalized Nash equilibria", which in particular does not imply convergence of learning to an equilibrium policy. In the next subsection, we propose a simple variation of Q-learning which converges to an equilibrium policy in all weakly acyclic stochastic games.

B. Q-Learning in Stochastic Dynamic Games

The discussion in the previous subsection reveals that the standard Q-learning (4)-(5) can lead to robust oscillations even in repeated coordination games. The main obstacle to convergence of Q-learning in games is due to the presence of multiple active learners leading to a non-stationary environment for all learners. To overcome this obstacle, we use some inspiration from our previous work [37] on repeated games and modify the Q-learning for stochastic games as follows. In our variation of Q-learning, we allow DMs to use constant policies for extended periods of time called *exploration phases*.

As illustrated in Figure 3, the $k-$th exploration phase runs through times $t = t_k, \ldots, t_{k+1} - 1$, where

$$t_{k+1} = t_k + T_k \qquad \text{(with } t_k = 0\text{)}$$

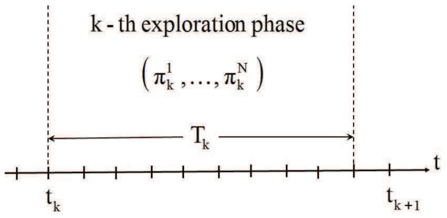

Fig. 3. An illustration of the $k-$th exploration phase.

for some integer $T_k \in [1, \infty)$ denoting the length of the $k-$th exploration phase. During the $k-$th exploration phase, DMs use some constant policies $\pi_k^1, \ldots, \pi_k^N$ as their baseline policies with occasional experimentation. The essence of the main idea is to create a stationary environment over each exploration phase so that DMs can accurately learn their optimal Q-factors corresponding to the constant policies used during each exploration phase. Before arguing why this would lead to an equilibrium policy in all weakly acyclic stochastic games, let us introduce our variation of Q-learning more precisely.

Algorithm 1 (for DM^i):

```
Set parameters
    $\mathbb{Q}^i$: some compact subset of the Euclidian space $\mathbb{R}^{|\mathbb{X}\times\mathbb{U}^i|}$
        where $|\mathbb{X}\times\mathbb{U}^i|$ is the number of pairs $(x, u^i)$
    $\{T_k\}_{k\geq 0}$: sequence of integers in $[1, \infty)$
    $\rho^i \in (0,1)$: experimentation probability
    $\lambda^i \in (0,1)$: inertia
    $\delta^i \in (0,\infty)$: tolerance level for sub-optimality
    $\{\alpha_n^i\}_{n\geq 0}$: sequence of step sizes where
            $\alpha_n^i \in [0,1]$, $\sum_n \alpha_n^i = \infty$, $\sum_n \left(\alpha_n^i\right)^2 < \infty$
            (e.g., $\alpha_n^i = 1/n^r$ where $r \in (1/2, 1]$)
Initialize $\pi_0^i \in \Pi^i$ (arbitrary), $Q_0^i \in \mathbb{Q}^i$ (arbitrary)
Receive $x_0$
Iterate $k \geq 0$
    ($k-$th exploration phase)
    Iterate $t = t_k, \ldots, t_{k+1}-1$
        $u_t^i = \begin{cases} \pi_k^i(x_t), & \text{w.p. } 1-\rho^i \\ \text{any } u^i \in \mathbb{U}^i, & \text{w.p. } \rho^i/|\mathbb{U}^i| \end{cases}$
        Receive $c^i(x_t, u_t^i, u_t^{-i})$
        Receive $x_{t+1}$ (selected according to $P[\,\cdot \mid x_t, u_t^i, u_t^{-i}]$)
        $n_t^i =$ the number of visits to $(x_t, u_t^i)$ in the $k-$th
                exploration phase up to $t$
        $Q_{t+1}^i(x_t, u_t^i) = (1-\alpha_{n_t^i}^i) Q_t^i(x_t, u_t^i)$
                $+\alpha_{n_t^i}^i \left[c^i(x_t, u_t^i, u_t^{-i}) + \beta^i \min_{v^i} Q_t^i(x_{t+1}, v^i)\right]$
        $Q_{t+1}^i(x, u^i) = Q_t^i(x, u^i)$, for all $(x, u^i) \neq (x_t, u_t^i)$
    End
    $\Pi_{k+1}^i = \big\{\hat{\pi}^i \in \Pi^i : Q_{t_{k+1}}^i(x, \hat{\pi}^i(x))$
                    $\leq \min_{v^i} Q_{t_{k+1}}^i(x, v^i) + \delta^i, \text{ for all } x\big\}$
    If $\pi_k^i \in \Pi_{k+1}^i$
        $\pi_{k+1}^i = \pi_k^i$
    Else
        $\pi_{k+1}^i = \begin{cases} \pi_k^i, & \text{w.p. } \lambda^i \\ \text{any } \pi^i \in \Pi_{k+1}^i, & \text{w.p. } (1-\lambda^i)/|\Pi_{k+1}^i| \end{cases}$
    End
    Reset $Q_{t_{k+1}}^i$ to any $Q^i \in \mathbb{Q}^i$ (e.g., project $Q_{t_{k+1}}^i$ onto $\mathbb{Q}^i$)
End
```

Algorithm 1 mimics the best reply process with inertia in §II-C arbitrarily closely with arbitrarily high probability under certain conditions. The key difference here is that each DM using Algorithm 1 approximately learns its optimal Q-factors during each exploration phase with limited observations. Accordingly, each DM updates its (baseline) policy to one of its near best replies with inertia based on its learnt Q-factors. Hence, Algorithm 1 can be regarded as an approximation to the best reply process with inertia in §II-C; see [41] where best replies are obtained based on rewards that must be estimated using noisy observations.

Assumption 1: For all (x', x), there exists a finite integer $H \geq 0$ and joint decisions $\tilde{u}_0, \ldots, \tilde{u}_H$ such that

$$P[x_{H+1} = x' \mid (x_0, u_0, \ldots, u_H) = (x, \tilde{u}_0, \ldots, \tilde{u}_H)] > 0.$$

Assumption 1 ensures that the step sizes satisfy the well-known conditions of the stochastic approximation theory [31] during each exploration phase.

Assumption 2: For all i, $0 < \delta^i < \bar{\delta}$ and $0 < \rho^i < \bar{\rho}$, where $\bar{\delta}$ and $\bar{\rho}$ (which depend only on the parameters of the game at hand) are defined in Appendix B.

Assumption 2 requires that the tolerance levels for sub-optimality used in the computation of near best replies as well as the experimentation probabilities be nonzero but sufficiently small.

Theorem 1: Consider a discounted stochastic game that is weakly acyclic under strict best replies. Suppose that each DM^i updates its policies by Algorithm 1. Let Assumption 1 and 2 hold.

(i) For any $\epsilon > 0$, there exist $\tilde{T} < \infty$, $\tilde{k} < \infty$ such that if $\min_\ell T_\ell \geq \tilde{T}$, then

$$P\left[\pi_k \in \Pi_{\text{eq}}\right] \geq 1 - \epsilon, \qquad \text{for all } k \geq \tilde{k}.$$

(ii) If $T_k \to \infty$, then

$$P\left[\pi_k \in \Pi_{\text{eq}}\right] \to 1.$$

(iii) There exists finite integers $\{\tilde{T}_k\}_{k\geq 0}$ such that if $T_k \geq \tilde{T}_k$, for all k, then

$$P\big[\pi_k \to \pi^*, \text{ for some } \pi^* \in \Pi_{\text{eq}}\big] = 1.$$

Proof: See Appendix B. ■

Let us discuss the main idea behind this result. Since all DMs use constant policies throughout any particular exploration phase, each DM indeed faces a stationary Markov decision problem in each exploration phase. Therefore, if the length of each exploration phase is long enough and the experimentation probabilities $\rho^1, \ldots, \rho^N$ are small enough (but non-zero), each DM^i can learn its corresponding optimal Q-factors in each exploration phase with arbitrary accuracy with arbitrarily high probability. This allows each DM^i to accurately compute its near best replies to the other DMs' policies π_k^{-i} at the end of the $k-$th exploration phase. Intuitively, allowing each DM^i to update its policy π_k^i to its near best replies (to π_k^{-i}) at the end of the $k-$th exploration phase with some inertia $\lambda^i \in (0,1)$ results in a policy adjustment process that approximates the best reply process with inertia in §II-C.

Remark 1: One may also wish to find explicit lower-bounds on T_k to achieve almost sure convergence based on the convergence rates of the standard Q-learning with a single DM; we refer the reader to [42] for bounds on the convergence rates for standard Q-learning.

IV. Generalizations

A. Learning in Weakly Acyclic Games under Strict Better Replies

We present another Q-learning algorithm with provable convergence to equilibrium in discounted stochastic games that are weakly acyclic under strict better replies. For this, we first introduce the notion of weak acyclicity under strict better replies. Given any $\pi = (\pi^i, \pi^{-i}) \in \Delta$, let Υ^i_π denote DM^i's set of (deterministic) better replies with respect to π, i.e.,

$$\Upsilon^i_\pi := \big\{\hat{\pi}^i \in \Pi^i : J_x(\hat{\pi}^i, \pi^{-i}) \leq J_x(\pi^i, \pi^{-i}), \text{ for all } x\big\}.$$

Any better reply $\hat{\pi}^i \in \Upsilon^i_\pi$ of DM^i is called a *strict better reply* (with respect to π) if

$$J^i_x(\hat{\pi}^i, \pi^{-i}) < J^i_x(\pi^i, \pi^{-i}), \quad \text{for some } x.$$

Definition 4: We call a (possibly finite) sequence of deterministic joint policies $\pi_0, \pi_1, \ldots$ a *strict better reply path* if, for each k, π_k and π_{k+1} differ in exactly one DM position, say DM^i, and π^i_{k+1} is a strict better reply with respect to π_k.

Definition 5: A discounted stochastic game is called weakly acyclic under strict better replies if there is a strict better reply path starting from each deterministic joint policy and ending at a deterministic equilibrium policy.

Since every strict best reply path is also a strict better reply path, the set of games weakly acyclic under better replies contain (in fact, strictly) the set of games weakly acyclic under best replies.

It is straightforward to introduce a policy adjustment process analogous to the one in §II-C where, at each step, each DM^i switches to one of its strict better replies with some inertia; see Sections 6.4-6.5 in [30]. Such a process would clearly converge to an equilibrium in games that are weakly acyclic under strict better replies. We next introduce a learning algorithm which allows each DM to learn the Q-factors corresponding to two policies, a baseline policy and a randomly selected experimental policy, during each exploration phase. If the learnt Q-factors indicate that the experimental policy is better than the baseline policy within a certain tolerance level, then the baseline policy is updated to the experimental policy with some inertia at the end of each exploration phase. This learning algorithm enables DMs to adjust their policies with much less information (as in §III-A), and follow (approximately) along the strict better reply paths that the adjustment process follows.

Algorithm 2 (for DM^i):

```
Set parameters as in Algorithm 1
Initialize $\pi^i_0, \hat{\pi}^i_0 \in \Pi^i$ (arbitrary except $\hat{\pi}^i_0 \neq \pi^i_0$), $Q^i_0, \hat{Q}^i_0 \in \mathbb{Q}^i$ (arbitrary)
Receive $x_0$
Iterate $k \geq 0$
    ($k-$th exploration phase)
    Iterate $t = t_k, \ldots, t_{k+1} - 1$
        $u^i_t = \begin{cases} \pi^i_k(x_t), & \text{w.p. } 1 - \rho^i \\ \text{any } u^i \in \mathbb{U}^i, & \text{w.p. } \rho^i/|\mathbb{U}^i| \end{cases}$
        Receive $c^i(x_t, u^i_t, u^{-i}_t)$
        Receive $x_{t+1}$ (selected according to $P[\,\cdot \mid x_t, u^i_t, u^{-i}_t]$)
        $n^i_t =$ the number of visits to $(x_t, u^i_t)$ in the $k-$th
                exploration phase up to $t$
        $Q^i_{t+1}(x_t, u^i_t) = (1 - \alpha^i_{n^i_t})Q^i_t(x_t, u^i_t)$
                $+\alpha^i_{n^i_t}\big[c^i(x_t, u^i_t, u^{-i}_t) + \beta^i Q^i_t(x_{t+1}, \pi^i_k(x_{t+1}))\big]$
        $\hat{Q}^i_{t+1}(x_t, u^i_t) = (1 - \alpha^i_{n^i_t})\hat{Q}^i_t(x_t, u^i_t)$
                $+\alpha^i_{n^i_t}\big[c^i(x_t, u^i_t, u^{-i}_t) + \beta^i Q^i_t(x_{t+1}, \hat{\pi}^i_k(x_{t+1}))\big]$
        $Q^i_{t+1}(x, u^i) = Q^i_t(x, u^i)$, for all $(x, u^i) \neq (x_t, u^i_t)$
        $\hat{Q}^i_{t+1}(x, u^i) = \hat{Q}^i_t(x, u^i)$, for all $(x, u^i) \neq (x_t, u^i_t)$
    End
    If $(\hat{Q}^i_{t_{k+1}}(x, \hat{\pi}^i_k(x)) \leq Q^i_{t_{k+1}}(x, \pi^i_k(x)) + \delta^i,$ for all $x)$
                                and
      $(\hat{Q}^i_{t_{k+1}}(x, \hat{\pi}^i_k(x)) \leq Q^i_{t_{k+1}}(x, \pi^i_k(x)) - \delta^i,$ for some $x)$
        $\pi^i_{k+1} = \begin{cases} \pi^i_k, & \text{w.p. } \lambda^i \\ \hat{\pi}^i_k, & \text{w.p. } 1 - \lambda^i \end{cases}$
    Else
        $\pi^i_{k+1} = \pi^i_k$
    End
    $\hat{\pi}^i_{k+1} =$ any policy $\pi^i \in \Pi^i \backslash \{\pi^i_{k+1}\}$ with equal
            probability
    Reset $Q^i_{t_{k+1}}$, $\hat{Q}^i_{t_{k+1}}$ to any $Q^i, \hat{Q}^i \in \mathbb{Q}^i$
End
```

Since any policy except the baseline policy can be chosen as an experimental policy (with equal probability), each DM can switch to any of its strict better replies with positive probability. In contrast, each DM using Algorithm 1 can only switch to one of its strict best replies. As a result, each DM using Algorithm 2 can escape a strict best reply cycle by switching to a strict better reply (if one exists); whereas, any DM using Algorithm 1 cannot. This flexibility comes at the cost of running two Q-learning recursions, one for the baseline policy and the other for the experimental policy, instead of one. However, this flexibility also leads to convergent behavior in a strictly larger set of games. We cite [43] as a reference to an earlier use of the idea of comparing two strategies and selecting one according to the Boltzman distribution.

The counterpart of Theorem 1 can be obtained for Algorithm 2 in games that are weakly acyclic under strict better replies.

Assumption 3: For all i, $0 < \delta^i < \check{\delta}$ and $0 < \rho^i < \check{\rho}$, where $\check{\delta}$ and $\check{\rho}$ (which depend only on the parameters of the game at hand) are defined in Appendix C.

Theorem 2: Consider a discounted stochastic game that is weakly acyclic under strict better replies. Suppose that each DM^i updates its policies by Algorithm 2. Let Assumption 1 and 3 hold.

(i) For any $\epsilon > 0$, there exist $\tilde{T} < \infty$, $\tilde{k} < \infty$ such that if $\min_\ell T_\ell \geq \tilde{T}$, then

$$P\left[\pi_k \in \Pi_{\text{eq}}\right] \geq 1 - \epsilon, \qquad k \geq \tilde{k}.$$

(ii) If $T_k \to \infty$, then

$$P\left[\pi_k \in \Pi_{\text{eq}}\right] \to 1.$$

(iii) There exists finite integers $\{\tilde{T}_k\}_{k\geq 0}$ such that if $T_k \geq \tilde{T}_k$, for all k, then

$$P\big[\pi_k \to \pi^*, \text{ for some } \pi^* \in \Pi_{\text{eq}}\big] = 1.$$

Proof: See Appendix C. ■

B. Learning in Weakly Acyclic Games under multi-DM Strict Best or Better Replies

The notion of weak acyclicity can be generalized by allowing multiple DMs to simultaneously update their policies in a strict best or better reply path.

Definition 6: We call a (possibly finite) sequence of deterministic joint policies $\pi_0, \pi_1, \ldots$ a *multi-DM strict best (better) reply path* if, for each k, π_k and π_{k+1} differ for at least one DM and, for each deviating DM^i, π^i_{k+1} is a strict best (better) reply with respect to π_k.

Definition 7: A discounted stochastic game is called weakly acyclic under multi-DM strict best (better) replies if there is a multi-DM strict best (better) reply path starting from each deterministic joint policy and ending at a deterministic equilibrium policy.

This generalization leads to a strictly larger set of games that are weakly acyclic. To see this, consider a single-stage game characterized by the cost matrices in Figure 4 where DM^1 chooses a row, DM^2 chooses a column, and DM^3 chooses a matrix, simultaneously. Assume $a > 0$. There is no strict best (or better) reply path to an equilibrium from the joint decisions $(1,1,1)$, $(1,3,1)$, $(3,3,1)$, $(3,1,1)$, $(1,1,2)$, $(3,1,2)$, if only a single DM can update its decision at a time. Therefore, this game is not weakly acyclic under strict best (or better) replies in the sense of Definition 3 (or Definition 5). However, if multiple DMs are allowed to switch to their strict best (or better) replies simultaneously, then it becomes possible to reach the equilibrium $(2,3,2)$ from any joint decision. For example, if DM^2 and DM^3 switch to their strict best (or better) replies simultaneously from the joint decision $(1,1,1)$, then the resulting joint decision would be $(1,3,2)$. This would subsequently lead to the equilibrium $(2,3,2)$ if DM^1 switches to its strict best (or better) reply from $(1,3,2)$.

	1	2	3
1	$-a,0,0$	$0,a,0$	$0,-a,-a$
2	$a,0,0$	$-a,-a,0$	$a,0,0$
3	$0,-a,-a$	$0,a,0$	$-a,0,-a$

1

	1	2	3
1	$0,-a,-a$	$0,0,0$	$0,0,0$
2	$a,0,0$	$-a,0,-a$	$-a,-a,-a$
3	$-a,-a,0$	$0,0,0$	$0,0,0$

2

Fig. 4. Cost matrices of a single-stage game with three DMs.

All learning algorithms introduced in the paper allow multiple DMs to simultaneously update their policies with positive probability. In view of this, it is straightforward to see that our main convergence results Theorem 1 (Theorem 2) hold in games that are weakly acyclic under multi-DM strict best (better) replies.

V. A Simulation Study: Prisoner's Dilemma with a State

We consider a discounted stochastic game with two DMs where $\mathbb{X} = \mathbb{U}^1 = \mathbb{U}^2 = \{1,2\}$. Each DM^i's utility (to be maximized) at each time $t \geq 0$ depends only on the joint decisions (u^1_t, u^2_t) of both DMs as

		DM^{-i}:	
		1	2
DM^i:	1	c	a
	2	b	0

Fig. 5. DM^i's single-stage utility.

We assume $b > c > 0 > a$. The state evolves as

$$P\big[x_{t+1} = 1 \mid (u^1_t, u^2_t) = (1,1)\big] = 1 - \gamma$$
$$P\big[x_{t+1} = 2 \mid (u^1_t, u^2_t) \neq (1,1)\big] = 1 - \gamma$$

where $\gamma \in (0,1)$ and $P[x_0 = 1] = 1/2$.

The single-stage game corresponds to the well-known prisoner's dilemma where the $i-$th prisoner (DM^i) cooperates (defects) at time t by choosing $u^i_t = 1$ ($u^i_t = 2$). The single-stage game has a unique equilibrium $(u^1, u^2) = (2,2)$, i.e., both DMs defect, leading to rewards $(0,0)$. The dilemma is that each DM can do strictly better by cooperating, i.e., $(u^1, u^2) = (1,1)$ (not an equilibrium).

In the multi-stage game, the state x_t indicates, w.p. $1-\gamma$, whether or not both DMs cooperated in the previous stage. It turns out that cooperation can be obtained as an equilibrium of the multi-stage game if the DMs are patient, i.e., the discount factors are sufficiently high, and the error probability γ is sufficiently small . Note that each DM^i has four different policies of the form $\pi^i : \mathbb{X} \to \mathbb{U}^i$. For large enough β^1, β^2, and small enough γ, the multi-stage game has two (Markov perfect) equilibria. In one equilibrium, called the cooperation equilibrium, each DM cooperates if $x = 1$ and defects otherwise. In the other equilibrium, called the defection equilibrium, both DMs always defect. Furthermore, from any joint policy in $\Pi^1 \times \Pi^2$, there is a strict best reply path to one of these two equilibria, which implies that the multi-stage game is weakly acyclic under strict best replies.

We set $b = 2$, $c = 1$, $a = -1$, $\gamma = 0.3$. We simulate Algorithm 1 with the following parameter values: $\rho^i = 0.1$, $\lambda^i = 0.5$, $\delta^i = 0$, $\alpha^i_k = 1/k^{0.51}$, for all i, k. We keep the lengths of the exploration phases constants, i.e, $T_k = T$, for all k. We consider different values for T since the lengths of the exploration phases appear to be most critical for the behavior of the learning process. For each value of T, we run Algorithm 1 and the best reply process with inertia (in §II-C) in parallel, with 1000 policy updates starting from each of the 16 initial joint policies in Π. We initialize all the learnt Q-factors at 0 for each simulation run; however, we do not reset the learnt Q-factors at the end of any exploration phase during any simulation run. We let π_k and $\breve{\pi}_k$ denote the policies generated by Algorithm 1 and the best reply process with inertia in §II-C, respectively. For each value of T, Table V shows the fraction of times at which π_k visits an equilibrium and the fraction of times at which π_k agrees with $\breve{\pi}_k$, during the 1000 policy updates (averaged uniformly over all 16 initial policies in Π).

The results in Table V reveals that, as T increases, π_k visits an equilibrium and agrees with $\breve{\pi}_k$ more often. This is consistent with Theorem 1 since DMs are expected to learn their Q-factors more accurately with higher probability for larger values of T. When T is sufficiently large, the polices π_k are at equilibrium and agrees with $\breve{\pi}_k$ nearly all of the time regardless of the initial policy. In a typical simulation run (with a large enough T), the polices π_k and $\breve{\pi}_k$ transition to an equilibrium in few steps and stay at equilibrium thereafter.

T	$\frac{1}{1001}\sum_{k=0}^{1000} I_{\{\pi_k\in\Pi_e\}}$ (averaged over $\pi_0 \in \Pi$)	$\frac{1}{1001}\sum_{k=0}^{1000} I_{\{\pi_k=\breve{\pi}_k\}}$ (averaged over $\pi_0 \in \Pi$)
10	0.2581	0.1254
25	0.5274	0.3410
50	0.7835	0.6170
100	0.9282	0.6301
1000	0.9935	0.6879
10000	0.9978	0.7733
50000	0.9976	0.9705

TABLE I
THE FRACTION OF TIMES AT WHICH π_k VISITS AN EQUILIBRIUM AND THE FRACTION OF TIMES AT WHICH π_k AGREES WITH $\breve{\pi}_k$.

VI. CONCLUDING REMARKS

In this paper, we develop decentralized Q-learning algorithms and present their convergence properties for stochastic games under weak acyclicity. This is the first paper, to our knowledge, that presents learning algorithms with convergence to equilibria in large classes of stochastic games. The decision makers observe only their own decisions and cost realizations, and the state transitions; they need not even know the presence of the other decision makers.

Our approach has a two-time scale flavor; however, unlike the existing work on multi-time-scale learning, it does not depend on the stochastic approximation theory. Note that the existing work on multi-time-scale learning, e.g., [11], [14], [32], [38], require the stability analysis of some ordinary differential equations (ODE) describing the mean behavior of the learning algorithms. Aside from the difficulty of choosing the step sizes running at multiple time scales, the existing work involves nonlinear ODEs whose analysis does not seem to be within reach even for dynamic team problems. In contrast, our approach leads to a considerably simpler analysis for all weakly acyclic stochastic games.

APPENDIX A
A UNIFORM CONVERGENCE RESULT FOR THE STANDARD Q-LEARNING ALGORITHM WITH A SINGLE DM

Convergence of the standard Q-learning algorithm with a single DM is well known [31]. However, to prove the results of this paper, we need the sample paths generated by the standard Q-learning algorithm to well behave with respect to the initial conditions. Let us now consider a single-DM version of the setup introduced in §II where the DM index i (in the superscript) is dropped (only in Appendix A) and $c(x,u)$ representing the one-stage cost for applying control u at x is an exogenous random variable with finite variance. Let us assume that a single DM using a stationary random policy $\pi \in \Delta$ updates its Q-factors as: for $t \geq 0$,

$$Q_{t+1}(x,u) = Q_t(x,u), \qquad \text{for all } (x,u) \neq (x_t,u_t) \quad (6)$$

$$Q_{t+1}(x_t,u_t) = Q_t(x_t,u_t) + \alpha_{n_t}\Big(c(x_t,u_t) + \beta \min_v Q_t(x_{t+1},v) - Q_t(x_t,u_t)\Big) \quad (7)$$

where the initial condition Q_0 is given, u_t is chosen according to $\pi(x_t)$, the state x_t evolves according to $P[\,\cdot\,|x_t,u_t]$ starting at x_0, n_t is the number of visits to (x_t,u_t) up to time t, and $\{\alpha_n\}_{n\geq 0}$ is a sequence of step sizes satisfying

$$\alpha_n \in [0,1], \qquad \sum_n \alpha_n = \infty, \qquad \sum_n \alpha_n^2 < \infty.$$

Lemma 1: Assume that each (x,u) is visited infinitely often w.p. 1. For any $\epsilon > 0$ and compact $\mathbb{Q} \in \mathbb{R}^{|\mathbb{X}\times\mathbb{U}|}$, there exists $T_\epsilon^{\mathbb{Q}} < \infty$ such that, for any $Q_0 \in \mathbb{Q}$,

$$P\left[\sup_{t\geq T_\epsilon^{\mathbb{Q}}} \left|Q_t - \bar{Q}\right|_\infty \leq \epsilon\right] \geq 1-\epsilon$$

where $|\cdot|_\infty$ denotes the maximum norm and $\bar{Q}$ is the unique fixed point of the mapping $F : \mathbb{X}\times\mathbb{U} \mapsto \mathbb{X}\times\mathbb{U}$ defined by

$$F(Q)(x,u) = E[c(x,u)] + \beta \sum_{x'} P[x'|x,u] \min_v Q(x',v)$$

for all x, u.

Proof: Let $\{Q'_t\}_{t\geq 0}$ and $\{Q''_t\}_{t\geq 0}$ be the trajectories for the initial conditions Q'_0 and Q''_0, respectively, corresponding to a sample path $\{(x_t,u_t,c(x_t,u_t))\}_{t\geq 0}$. It is easy to see that, for all $t \geq 0$,

$$|Q'_{t+1}(x_t,u_t) - Q''_{t+1}(x_t,u_t)| \leq (1-\alpha_{n_t})|Q'_t(x_t,u_t) - Q''_t(x_t,u_t)| + \alpha_{n_t}\beta|Q'_t - Q''_t|_\infty.$$

This implies that $M_t := \sup_{Q'_0,Q''_0\in\mathbb{Q}} |Q'_t - Q''_t|_\infty$ is non-increasing and therefore convergent. Suppose that $M_t \to \bar{M} > 0$. There exists some $\bar{t} < \infty$ such that $\max_{t\geq\bar{t}} M_t < \bar{M}(1+1/\beta)/2$. Hence, we have, for all $t \geq \bar{t}$,

$$|Q'_{t+1}(x_t,u_t) - Q''_{t+1}(x_t,u_t)| \leq (1-\alpha_{n_t})|Q'_t(x_t,u_t) - Q''_t(x_t,u_t)| + \alpha_{n_t}\beta\frac{\bar{M}(1+1/\beta)}{2}.$$

This leads to: for all (x,u) and $t \geq \bar{t}$,

$$|Q'_{t+1}(x,u) - Q''_{t+1}(x,u)| \leq \Pi_{s=0}^{m_t(x,u)}(1-\alpha_s)M_0 + \left[1 - \Pi_{s=0}^{m_t(x,u)}(1-\alpha_s)\right]\beta\bar{M}(1+1/\beta)/2$$

where $m_t(x,u) := \sum_{s=0}^t I_{\{(x_t,u_t)=(x,u)\}}$ is the number of visits to (x,u) in $[0,t]$. Since each (x,u) is visited infinitely often w.p. 1 and $\sum_s \alpha_s = \infty$, we have, for each (x,u), $\Pi_{s=0}^{m_t(x,u)}(1-\alpha_s) \to 0$ as $t\to\infty$ w.p. 1. This implies that $\bar{M} \leq \beta\bar{M}(1+1/\beta)/2 < \bar{M}$ w.p. 1, which is a contradiction. Therefore, $M_t \to 0$, w.p. 1.

Theorem 4 in [31] shows that, for any initial condition Q_0, $Q_t \to \bar{Q}$, w.p. 1. Hence, for any $Q'_0 \in \mathbb{Q}$, we have $|Q'_t - \bar{Q}|_\infty + \sup_{Q''_0\in\mathbb{Q}} |Q'_t - Q''_t|_\infty \to 0$, w.p. 1. Therefore, $\sup_{Q''_0\in\mathbb{Q}} |Q''_t -$

$\bar{Q}|_\infty \to 0$, w.p. 1. This leads to the desired result, i.e., for any $\epsilon > 0$ and compact $\mathbb{Q} \in \mathbb{R}^{|\mathbb{X}\times\mathbb{U}|}$, there exists $T_\epsilon^{\mathbb{Q}} < \infty$ such that

$$P\left[\sup_{t\geq T_\epsilon^{\mathbb{Q}}} \sup_{Q_0''\in\mathbb{Q}} |Q_t'' - \bar{Q}|_\infty \leq \epsilon\right] \geq 1-\epsilon.$$

■

Remark 2: The Q-factors corresponding to a certain deterministic policy $\hat{\pi}$ can be learnt by modifying the recursion (6)-(7) as follows: for $t \geq 0$,

$$\begin{aligned} \hat{Q}_{t+1}(x,u) &= \hat{Q}_t(x,u), \qquad \text{for all } (x,u) \neq (x_t,u_t) \\ \hat{Q}_{t+1}(x_t,u_t) &= \hat{Q}_t(x_t,u_t) + \alpha_{n_t}\big(c(x_t,u_t) \\ &\qquad + \beta\hat{Q}_t(x_{t+1},\hat{\pi}(x_{t+1})) - \hat{Q}_t(x_t,u_t)\big) \end{aligned}$$

where the initial condition $\hat{Q}_0$ is given and u_t is chosen according to $\pi(x_t)$. Hence, the uniform convergence result in Lemma 1 also holds for the this recursion.

Appendix B
Proof of Theorem 1

For any $\pi^{-i} \in \Delta^{-i}$, let $F^i_{\pi^{-i}}$ denote the self-mapping of $\mathbb{X}\times\mathbb{U}^i$ defined by

$$\begin{aligned} F^i_{\pi^{-i}}(Q^i)(x,u^i) =& E_{\pi^{-i}(x)}\big[c^i\left(x,u^i,u^{-i}\right) \\ &+ \beta^i \sum_{x'} P\left[x'|x,u^i,u^{-i}\right] \min_{v^i} Q^i(x',v^i)\big] \end{aligned}$$

for all x, u^i. It is well-known that $F^i_{\pi^{-i}}$ is a contraction mapping with the Lipschitz constant β^i with respect to the maximum norm. Recall from (2) that each DM^i's optimal Q-factors $Q^i_{\pi^{-i}}$ is the unique fixed point of $F^i_{\pi^{-i}}$. We also note that, during the $k-$th exploration phase, each DM^i actually uses the random policy $\bar{\pi}^i_k$ defined as

$$\bar{\pi}^j_k = (1-\rho^j)\pi^j_k + \rho^j\nu^j \tag{8}$$

where ν^j is the random policy that assigns the uniform distribution on $\mathbb{U}^j$ to each x.

Lemma 2: For any $\epsilon > 0$, there exists $T_\epsilon < \infty$ such that, if $T_k \geq T_\epsilon$, then

$$P\left[\left|Q^i_{t_{k+1}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty \leq \epsilon, \text{ for all } i\right] \geq 1-\epsilon.$$

Proof: Note that the $k-$th exploration phase starts with x_{kT}, which belongs to the finite state space $\mathbb{X}$, and $Q^i_{t_k} \in \mathbb{Q}^i$, where $\mathbb{Q}^i$ is compact, for all i. Note also that, during each exploration phase, DMs use stationary random policies of the form (8) and there are finitely many such joint policies. Hence, the desired result follows from Lemma 1 in Appendix A. ■

Lemma 3: For any $\epsilon > 0$, there exists $\rho_\epsilon > 0$ such that, if $\rho^i \leq \rho_\epsilon$, for all i, then

$$\left|Q^i_{\pi_k^{-i}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty \leq \epsilon, \qquad \text{for all } i,k.$$

Proof: We have

$$\begin{aligned} \left|Q^i_{\pi_k^{-i}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty &= \left|F^i_{\pi_k^{-i}}(Q^i_{\pi_k^{-i}}) - F^i_{\bar{\pi}_k^{-i}}(Q^i_{\bar{\pi}_k^{-i}})\right|_\infty \\ &\leq \left|F^i_{\pi_k^{-i}}(Q^i_{\pi_k^{-i}}) - F^i_{\bar{\pi}_k^{-i}}(Q^i_{\pi_k^{-i}})\right|_\infty \\ &\quad + \left|F^i_{\bar{\pi}_k^{-i}}(Q^i_{\pi_k^{-i}}) - F^i_{\bar{\pi}_k^{-i}}(Q^i_{\bar{\pi}_k^{-i}})\right|_\infty \\ &\leq \left(1 - \prod_{j\neq i}(1-\rho^j)\right) \times \\ &\quad \left|F^i_{\pi_k^{-i}}(Q^i_{\pi_k^{-i}}) - F^i_{\phi_k^{-i}}(Q^i_{\pi_k^{-i}})\right|_\infty \\ &\quad + \beta^i \left|Q^i_{\pi_k^{-i}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty \end{aligned}$$

where $\phi_k^{-i} \in \Delta^{-i}$ is some convex combination of the policies in Δ^{-i} of the form where each DM^j, $j \neq i$, either uses its baseline policy $\pi^j_k \in \Pi^j$ or the uniform distribution[3]. Because $(\pi_k^{-i}, \phi_k^{-i})$ belongs to a finite subset of $\Pi^{-i}\times\Delta^{-i}$, an upper bound $\bar{F} < \infty$ on

$$\left|F^i_{\pi_k^{-i}}(Q^i_{\pi_k^{-i}}) - F^i_{\phi_k^{-i}}(Q^i_{\pi_k^{-i}})\right|_\infty$$

exists, which is uniform in $(\pi_k^{-i}, \phi_k^{-i})$. This results in

$$\left|Q^i_{\pi_k^{-i}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty \leq \left(1 - \prod_{j\neq i}(1-\rho^j)\right)\frac{\bar{F}}{1-\beta^i}$$

which proves the lemma. ■

Let $\bar{\delta}$ denote the minimum separation between the entries of DMs' optimal Q-factors (with respect to the deterministic policies), defined as[4]

$$\bar{\delta} := \min_{\substack{i,x,v^i,\tilde{v}^i,\pi^{-i}\in\Pi^{-i}: \\ Q^i_{\pi^{-i}}(x,v^i)\neq Q^i_{\pi^{-i}}(x,\tilde{v}^i)}} \left|Q^i_{\pi^{-i}}(x,v^i) - Q^i_{\pi^{-i}}(x,\tilde{v}^i)\right|.$$

We consider $\bar{\delta}$ to be an upper bound on the tolerance levels for sub-optimality, i.e., $\delta^i \in (0,\bar{\delta})$, for all i. In that case, we also introduce an upper bound $\bar{\rho} > 0$ on the experimentation rates such that, if $\rho^i \leq \bar{\rho}$, for all i, then

$$\left|Q^i_{\pi_k^{-i}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty < \frac{1}{2}\min\{\delta^i, \bar{\delta}-\delta^i\}, \quad \text{for all } i,\ k. \tag{9}$$

Such an upper bound $\bar{\rho} > 0$ exists due to Lemma 3.

Lemma 4: Suppose $\delta^i \in (0,\bar{\delta})$, $\rho^i \in (0,\bar{\rho})$, for all i. For any $\epsilon > 0$, there exist $\bar{T} < \infty$, such that, if $T_k \geq \bar{T}$, then

$$P\left[E_k\right] \geq 1-\epsilon$$

where E_k, $k \geq 0$, is the random event defined as

$$\begin{aligned} E_k := \Big\{\omega\in\Omega : \left|Q^i_{t_{k+1}} - Q^i_{\pi_k}\right|_\infty <& \frac{1}{2}\min\{\delta^i, \bar{\delta}-\delta^i\}, \\ & \text{for all } i\Big\}. \end{aligned}$$

Proof: The desired result follows from Lemma 2 and (9). ■

[3] More precisely, $\phi_k^{-i} = \sum_{J\subset\{1,\ldots,N\}\backslash\{i\}} a_J\phi_{k,J}^{-i}$ where $a_J := \frac{\prod_{j\in J}(1-\rho^j)\prod_{j\notin J\cup\{i\}}\rho^j}{1-\prod_{j\neq i}(1-\rho^j)}$ and $\phi_{k,J} \in \Delta^{-i}$ is a policy such that $\phi_{k,J}^j = \pi_k^j$ for $j \in J$ and $\phi_{k,J}^j = \nu^j$ for $j \notin J\cup\{i\}$.

[4] To avoid trivial cases, we assume $Q^i_{\pi^{-i}}(x,v^i) \neq Q^i_{\pi^{-i}}(x,\tilde{v}^i)$ for some i, x, v^i, $\tilde{v}^i$, $\pi^{-i} \in \Pi^{-i}$.

A. Proof of part (i)

Note that

$$\omega \in E_k \Rightarrow \Pi_{\pi_k} = \Pi_{k+1} = \Pi^1_{k+1} \times \cdots \times \Pi^N_{k+1}.$$

Therefore, we have

$$P\left[\pi_{k+1} = \pi_k | E_k,\ \pi_k \in \Pi_{\text{eq}}\right] = 1, \quad \text{for all } k. \tag{10}$$

Since we have a weakly acyclic game at hand, for each $\pi \in \Pi$, there exists a strict best reply path of minimum length $L_\pi < \infty$ starting at π and ending at an equilibrium policy. Let $L := \max_{\pi\in\Pi} L_\pi$. There exists $p_{\min} \in (0,1)$ (which depends only on $\lambda^1, \ldots, \lambda^N$, and L) such that, for all k,

$$P\big[\pi_{k+L} \in \Pi_{\text{eq}} \big| E_k, \ldots, E_{k+L-1}, \pi_k \notin \Pi_{\text{eq}}\big] \geq p_{\min}. \tag{11}$$

Pick $\tilde{\epsilon} \in (0, \epsilon)$ satisfying

$$\left(\frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon} + (1-\tilde{\epsilon})p_{\min}} - \tilde{\epsilon}\right)(1-\tilde{\epsilon}) \geq 1 - \epsilon.$$

Lemma 4 implies the existence of $\tilde{T} < \infty$ such that, if $\min_\ell T_\ell \geq \tilde{T}$, then

$$P\left[E_k, \ldots, E_{k+L-1}\right] \geq 1 - \tilde{\epsilon}, \quad \text{for all } k. \tag{12}$$

For the rest of this part, we assume $\min_\ell T_\ell \geq \tilde{T}$. From (10), (11), (12), we obtain

$$P\left[\pi_{k+L} \in \Pi_{\text{eq}} | \pi_k \notin \Pi_{\text{eq}}\right] \geq p_{\min}(1-\tilde{\epsilon}), \quad \text{for all } k$$

and

$$P\left[\pi_{k+L} = \cdots = \pi_k | \pi_k \in \Pi_{\text{eq}}\right] \geq 1 - \tilde{\epsilon}, \quad \text{for all } k. \tag{13}$$

This leads to the recursive inequalities

$$p_{(n+1)L} \geq (1-\tilde{\epsilon})[p_{nL} + p_{\min}(1-p_{nL})] \tag{14}$$

where $p_k := P\left[\pi_k \in \Pi_{\text{eq}}\right]$, for all k. Note that we have, for all n,

$$p_{(n+1)L} - p_{nL} \geq -\tilde{\epsilon}. \tag{15}$$

We rewrite (14) as

$$p_{(n+1)L} - p_{nL} \geq [\tilde{\epsilon} + (1-\tilde{\epsilon})p_{\min}]\left[\frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon} + (1-\tilde{\epsilon})p_{\min}} - p_{nL}\right].$$

This shows that if

$$p_{nL} \leq \frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon} + (1-\tilde{\epsilon})p_{\min}} - \tilde{\epsilon} \tag{16}$$

we have $p_{(n+1)L} \geq p_{nL} + p_{\min}\tilde{\epsilon}$. Therefore, whenever p_{nL} satisfies (16), it will increase by at least $p_{\min}\tilde{\epsilon}$ until it exceeds the right hand side of (16), which will happen in a finite number of steps. In fact, p_{nL} would increase as long as $p_{nL} < \frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon}+(1-\tilde{\epsilon})p_{\min}}$. On the other hand, if $p_{nL} \geq \frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon}+(1-\tilde{\epsilon})p_{\min}}$, p_{nL} cannot decrease more than $\tilde{\epsilon}$; recall (15). Therefore, there exists $\tilde{n} < \infty$ such that, for all $n \geq \tilde{n}$,

$$p_{nL} \geq \frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon} + (1-\tilde{\epsilon})p_{\min}} - \tilde{\epsilon}.$$

Finally, due to (13), we have, for all $n \geq \tilde{n}$, $\ell \in \{1, \ldots, L-1\}$,

$$p_{nL+\ell} \geq \left(\frac{(1-\tilde{\epsilon})p_{\min}}{\tilde{\epsilon} + (1-\tilde{\epsilon})p_{\min}} - \tilde{\epsilon}\right)(1-\tilde{\epsilon}) \geq 1 - \epsilon.$$

B. Proof of part (ii)

For any $\epsilon > 0$, let $\tilde{T} < \infty$, $\tilde{k} < \infty$ be as in part (i). Let $\hat{k} < \infty$ be such that $\min_{k \geq \hat{k}} T_k \geq \tilde{T}$. It is straightforward to see from the proof of part (i) that, for all $k \geq \hat{k} + \tilde{k}$, we have $P\left[\pi_k \in \Pi_{\text{eq}}\right] \geq 1 - \epsilon$.

C. Proof of part (iii)

Pick a sequence $\{\tilde{\epsilon}_n\}_{n\geq 0}$ satisfying $\tilde{\epsilon}_n > 0$, for all n, and

$$\sum_n (1-p_{\min})^{-n}\tilde{\epsilon}_n < \infty \tag{17}$$

where $p_{\min}$ is as in (11). Lemma 4 implies the existence of a sequence $\{\tilde{T}_n\}_{n\geq 0}$ of finite integers such that if

$$T_{nL}, \ldots, T_{(n+1)L-1} \geq \tilde{T}_n \tag{18}$$

then

$$P\left[E_{nL}, \ldots, E_{(n+1)L-1}\right] \geq 1 - \tilde{\epsilon}_n. \tag{19}$$

We assume (18) (therefore (19)) holds for all n. This leads to

$$P[\pi_{(n+1)L} \notin \Pi_{\text{eq}}] \leq (1-p_{\min})P\left[\pi_{nL} \notin \Pi_{\text{eq}}\right] + \tilde{\epsilon}_n.$$

From this, it is straightforward to obtain

$$P\left[\pi_{(n+1)L} \notin \Pi_e\right] \leq (1-p_{\min})^n\left(1 + \sum_{s=0}^{n}(1-p_{\min})^{-s}\tilde{\epsilon}_s\right).$$

Due to (19), we have, for $\ell \in \{0, \ldots, L-1\}$,

$$P\left[\pi_{nL+\ell} \in \Pi_{\text{eq}}\right] \geq (1-\tilde{\epsilon}_n)P\left[\pi_{nL} \in \Pi_{\text{eq}}\right].$$

Therefore, for $\ell \in \{0, \ldots, L-1\}$,

$$P\left[\pi_{(n+1)L+\ell} \notin \Pi_{\text{eq}}\right] \leq (1-p_{\min})^n\left(1 + \sum_{s=0}^{n}(1-p_{\min})^{-s}\tilde{\epsilon}_s\right) + \tilde{\epsilon}_{n+1}.$$

From this and (17), we obtain

$$\sum_{k\geq 1} P\left[\pi_k \notin \Pi_e\right] \leq L\sum_{n\geq 0}\left[(1-p_{\min})^n\left(1 + \sum_{s=0}^{n}(1-p_{\min})^{-s}\tilde{\epsilon}_s\right) + \tilde{\epsilon}_{n+1}\right] < \infty.$$

Borel-Cantelli Lemma implies

$$P[\pi_k \notin \Pi_{\text{eq}},\ \text{for infinitely many } k] = 0. \tag{20}$$

From (17) and (19), we obtain $\sum_{k\geq 0} P\left[\Omega \backslash E_k\right] < \infty$. Borel-Cantelli Lemma again implies

$$P[\Omega \backslash E_k,\ \text{for infinitely many } k] = 0. \tag{21}$$

Finally, (20) and (21) imply the desired result.

Appendix C
Proof of Theorem 2

For any $\pi = (\pi^i, \pi^{-i}) \in \Pi^i \times \Delta^{-i}$, let F_π^i denote the self-mapping of $\mathbb{X} \times \mathbb{U}^i$ defined by

$$\begin{aligned} F_\pi^i(Q^i)(x,u^i) =& E_{\pi^{-i}(x)}\big[c^i\left(x,u^i,u^{-i}\right) \\ &+ \beta^i \sum_{x'} P\left[x'|x,u^i,u^{-i}\right] Q^i(x', \pi^i(x'))\big] \end{aligned}$$

for all x, u^i. It is well-known that F_π^i is a contraction mapping with the Lipschitz constant β^i with respect to the maximum norm. Let us denote the unique fixed point of F_π^i by Q_π^i. We also note that, during the $k-$th exploration phase, each DM^i actually uses the random policy $\bar{\pi}_k^i$ defined as

$$\bar{\pi}_k^j = (1-\rho^j)\pi_k^j + \rho^j \nu^j \tag{22}$$

where ν^j is the random policy that assigns the uniform distribution on $\mathbb{U}^j$ to each x.

Lemma 5: For any $\epsilon > 0$, there exists $T_\epsilon < \infty$ such that, if $T \geq T_\epsilon$, then

$$\begin{aligned} P\Big[& \left|Q_{t_{k+1}}^i - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty \leq \epsilon \text{ and} \\ & \left|\hat{Q}_{t_{k+1}}^i - Q^i_{(\hat{\pi}_k^i, \bar{\pi}_k^{-i})}\right|_\infty \leq \epsilon, \text{for all } i\Big] \geq 1-\epsilon, \text{ for all } k. \end{aligned}$$

Proof: Note that each exploration phase starts with x_{kT}, which belongs to a finite state space, and $Q_{kT}^i, \hat{Q}_{kT}^i \in \mathbb{Q}^i$, where $\mathbb{Q}^i$ is compact, for all i. Note also that, during each exploration phase, DMs use stationary random policies of the form (22) and there are finitely many such joint policies. Hence, the desired result follows from Lemma 1 in Appendix A; see Remark 2. ■

Lemma 6: For any $\epsilon > 0$, there exists $\rho_\epsilon > 0$ such that, if $\rho^i \leq \rho_\epsilon$, for all i, then

$$\begin{aligned} & \left|Q^i_{(\pi_k^i, \pi_k^{-i})} - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty \leq \epsilon \\ \text{and} & \left|Q^i_{(\hat{\pi}_k^i, \pi_k^{-i})} - Q^i_{(\hat{\pi}_k^i, \bar{\pi}_k^{-i})}\right|_\infty \leq \epsilon, \qquad \text{for all } i,k. \end{aligned}$$

Proof: We have

$$\begin{aligned} & \left|Q^i_{(\pi_k^i, \pi_k^{-i})} - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty \\ & = \left|F^i_{(\pi_k^i, \pi_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right) - F^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\left(Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right)\right|_\infty \\ & \leq \left|F^i_{(\pi_k^i, \pi_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right) - F^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right)\right|_\infty \\ & \quad + \left|F^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right) - F^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\left(Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right)\right|_\infty \\ & \leq \left(1 - \prod_{j\neq i}(1-\rho^j)\right) \times \\ & \qquad \left|F^i_{(\pi_k^i, \pi_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right) - F^i_{(\pi_k^i, \phi_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right)\right|_\infty \\ & \quad + \beta^i \left|Q^i_{(\pi_k^i, \pi_k^{-i})} - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty \end{aligned}$$

where $\phi_k^{-i} \in \Delta^{-i}$ is some convex combination of the joint policies of the form where each DM^j, $j \neq i$, either uses its baseline policy $\pi_k^j \in \Pi^j$ or the uniform distribution (as in Appendix B). Because $(\pi_k^i, \pi_k^{-i}, \phi_k^{-i})$ belongs to a finite subset of $\Pi^i \times \Pi^{-i} \times \Delta^{-i}$, an upper bound $\check{F} < \infty$ on

$$\left|F^i_{(\pi_k^i, \pi_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right) - F^i_{(\pi_k^i, \phi_k^{-i})}\left(Q^i_{(\pi_k^i, \pi_k^{-i})}\right)\right|_\infty$$

exists, which is uniform in $(\pi_k^i, \pi_k^{-i}, \phi_k^{-i})$. This results in

$$\left|Q^i_{(\pi_k^i, \pi_k^{-i})} - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty \leq \left(1 - \prod_{j\neq i}(1-\rho^j)\right) \frac{\check{F}}{1-\beta^i}$$

which leads to the first bound. The second bound can be obtained similarly. ■

Let $\check{\delta}$ denote the minimum separation between the entries of DMs' Q-factors (for deterministic policies), defined as[5]

$$\begin{aligned} \check{\delta} := \min\Big\{ & \left|Q^i_{(\pi^i, \pi^{-i})}(x, \pi^i(x)) - Q^i_{(\tilde{\pi}^i, \pi^{-i})}(x, \tilde{\pi}^i(x))\right| : \\ & i, x, \pi^i, \tilde{\pi}^i \in \Pi^i, \pi^{-i} \in \Pi^{-i}, \\ & Q^i_{(\pi^i, \pi^{-i})}(x, \pi^i(x)) \neq Q^i_{(\tilde{\pi}^i, \pi^{-i})}(x, \tilde{\pi}^i(x))\Big\}. \end{aligned}$$

We consider $\check{\delta}$ to be an upper bound on the tolerance levels for sub-optimality, i.e., $\delta^i \in (0, \check{\delta})$, for all i. In that case, we also introduce an upper bound $\check{\rho} > 0$ on the experimentation rates such that, if $\rho^i \leq \check{\rho}$, for all i, then

$$\begin{aligned} \max\Big\{ & \left|Q^i_{(\pi_k^i, \pi_k^{-i})} - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty, \\ & \left|Q^i_{(\hat{\pi}_k^i, \pi_k^{-i})} - Q^i_{(\hat{\pi}_k^i, \bar{\pi}_k^{-i})}\right|_\infty \Big\} < \frac{1}{2}\min\{\delta^i, \check{\delta} - \delta^i\} \end{aligned} \tag{23}$$

for all i, k. Such an upper bound $\check{\rho} > 0$ exists due to Lemma 6.

Lemma 7: Suppose $0 < \delta^i < \check{\delta}$, $0 < \rho^i < \check{\rho}$, for all i. For any $\epsilon > 0$, there exist $\bar{T} < \infty$, such that, if $T_k \geq \bar{T}$, then

$$P\left[\check{E}_k\right] \geq 1 - \epsilon$$

where $\check{E}_k$, $k \geq 0$, is the random event defined as

$$\begin{aligned} \check{E}_k := \Big\{\omega \in \Omega : \max\Big\{ & \left|Q^i_{t_{k+1}} - Q^i_{(\pi_k^i, \bar{\pi}_k^{-i})}\right|_\infty, \\ & \left|\hat{Q}^i_{t_{k+1}} - Q^i_{(\hat{\pi}_k^i, \bar{\pi}_k^{-i})}\right|_\infty \Big\} \\ & < \frac{1}{2}\min\{\delta^i, \check{\delta} - \delta^i\}, \text{ for all } i\Big\}. \end{aligned}$$

Proof: The desired result follows from Lemma 5 and (23). ■

We have

$$P\left[\pi_{k+1} = \pi_k | \check{E}_k,\ \pi_k \in \Pi_{\text{eq}}\right] = 1, \qquad \text{for all } k. \tag{24}$$

Since we have a weakly acyclic game at hand, for each $\pi \in \Pi$, there exists a strict better reply path of minimum length $\check{L}_\pi < \infty$ starting at π and ending at an equilibrium policy. Let $\check{L} := \max_{\pi \in \Pi} \check{L}_\pi$. There exists $\check{p}_{\min} \in (0,1)$ (which depends only on $\lambda^1, \ldots, \lambda^N$, and L) such that, for all k,

$$P\big[\pi_{k+\check{L}} \in \Pi_{\text{eq}} \big| \check{E}_k, \ldots, \check{E}_{k+L-1}, \pi_k \notin \Pi_{\text{eq}}\big] \geq \check{p}_{\min}. \tag{25}$$

Pick $\check{\epsilon} \in (0, \epsilon)$ satisfying

$$\left(\frac{(1-\check{\epsilon})\check{p}_{\min}}{\check{\epsilon} + (1-\check{\epsilon})\check{p}_{\min}} - \check{\epsilon}\right)(1-\check{\epsilon}) \geq 1 - \epsilon.$$

[5] We assume $Q^i_{(\pi^i, \pi^{-i})}(x, \pi^i(x)) \neq Q^i_{(\tilde{\pi}^i, \pi^{-i})}(x, \tilde{\pi}^i(x))$, for some i, x, $\pi^i, \tilde{\pi}^i \in \Pi^i$, $\pi^{-i} \in \Pi^{-i}$, to avoid trivial cases.

Lemma 7 implies the existence of $\check{T} < \infty$ such that, if $\min_\ell T_\ell \geq \check{T}$, then

$$P\left[\check{E}_k, \ldots, \check{E}_{k+L-1}\right] \geq 1 - \check{\epsilon}, \quad \text{for all } k. \tag{26}$$

For the rest of the proof, we assume $\min_\ell T_\ell \geq \check{T}$. From (24), (25), (26), we obtain, for all k,

$$\begin{aligned} &P\left[\pi_{k+\check{L}} \in \Pi_{\mathrm{eq}} | \pi_k \notin \Pi_{\mathrm{eq}}\right] \geq \check{p}_{\min}(1-\check{\epsilon}) \\ &\text{and } P\left[\pi_{k+\check{L}} = \cdots = \pi_k | \pi_k \in \Pi_{\mathrm{eq}}\right] \geq 1 - \check{\epsilon}. \end{aligned}$$

This leads to the recursive inequalities

$$p_{(n+1)\check{L}} \geq (1-\check{\epsilon})[p_{n\check{L}} + \check{p}_{\min}(1-p_{n\check{L}})], \quad n \geq 0 \tag{27}$$

where $p_k := P\left[\pi_k \in \Pi_{\mathrm{eq}}\right]$. Note that these inequalities are similar to (14) and by similar reasoning, there exists $\check{n} < \infty$ such that, for all $n \geq \check{n}$ and $\ell \in \{1, \ldots, L-1\}$,

$$p_{n\check{L}+\ell} \geq \left(\frac{(1-\check{\epsilon})\check{p}_{\min}}{\check{\epsilon} + (1-\check{\epsilon})\check{p}_{\min}} - \check{\epsilon}\right)(1-\check{\epsilon}) \geq 1 - \epsilon.$$

This proves part (i). The proofs of part (ii)-(iii) are analogous to the proofs of part (ii)-(iii) of Theorem 1, respectively.

References

[1] A. M. Fink *et al.*, "Equilibrium in a stochastic n-person game," *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, vol. 28, no. 1, pp. 89–93, 1964.

[2] C. J. C. H. Watkins and P. Dayan, "Q-Learning," *Machine Learning*, vol. **8**, no. 3, pp. 279–292, 1992.

[3] Y. Shoham, R. Powers, and T. Grenager, "If multi-agent learning is the answer, what is the question?" *Artificial Intelligence*, vol. 171, no. 7, pp. 365–377, 2007.

[4] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the Tenth International Conference on Machine Learning, Amherst, MA*, 1993, pp. 330–337.

[5] S. Sen, M. Sekaran, and J. Hale, "Learning to coordinate without sharing information," in *Proceedings of the 12th National Conference on Artificial Intelligence*, 1994, pp. 426–431.

[6] G. Brown, "Iterative solutions of games by fictitious play," in *Activity Analysis of Production and Allocation*, T. Koopmans, Ed. New York: Wiley, 1951, pp. 374–376.

[7] J. Robinson, "An iterative method of solving a game," *Annals of Mathematics*, vol. **54**, pp. 296–301, 1951.

[8] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Tenth Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin*, July 1998, pp. 746–752.

[9] G. Schoenmakers, J. Flesch, and F. Thuijsman, "Fictitious play in stochastic games," *Mathematical Methods of Operations Research*, vol. 66, no. 2, pp. 315–325, 2007.

[10] O. J. Vrieze and S. H. Tijs, "Fictitious play applied to sequence of games and discounted stochastic games," *International Journal of Game Theory*, vol. **11**, pp. 71–85, 1980.

[11] V. Borkar, "Reinforcement learning in Markovian evolutionary games," *Advances in Complex Systems*, vol. **5**, no. 1, pp. 55–72, 2002.

[12] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. **136**, pp. 215–250, 2002.

[13] ——, "Scalable learning in stochastic games," in *Proceedings of the AAAI Workshop on Game Theoretic and Decision Theoretic Agents, Edmonton, Canada*, 2002.

[14] V. R. Konda and V. S. Borkar, "Actor-critic-type learning algorithms for Markov decision processes," *SIAM Journal on Control and Optimization*, vol. **38**, no. 1, pp. 94–123, 1999.

[15] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press, 1961.

[16] R. Isaacs, *Differential Games*. New York, NY: Wiley, 1965.

[17] T. Başar and P. Bernhard, *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Boston, MA: Birkhäuser, 1995.

[18] E. Altman, "Flow control using the theory of zero sum markov games," *IEEE Transactions on Automatic Control*, vol. 39, no. 4, pp. 814–818, 1994.

[19] J. Ding, M. Kamgarpour, S. Summers, A. Abate, J. Lygeros, and C. Tomlin, "A stochastic games framework for verification and control of discrete time stochastic hybrid systems," *Automatica*, vol. 49, no. 9, pp. 2665–2674, 2013.

[20] Q. Zhu and T. Başar, "Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: Games-in-games principle for optimal cross-layer resilient control systems," *Control Systems, IEEE*, vol. 35, no. 1, pp. 46–65, 2015.

[21] E. Altman, "Non zero-sum stochastic games in admission, service and routing control in queueing systems," *Queueing Systems*, vol. 23, no. 1-4, pp. 259–279, 1996.

[22] J. W. Huang and V. Krishnamurthy, "Transmission control in cognitive radio as a Markovian dynamic game: Structural result on randomized threshold policies," *IEEE Transactions on Communications*, vol. 58, no. 1, pp. 301–310, 2010.

[23] Q. Zhu, H. Tembine, and T. Başar, "Network security configurations: A nonzero-sum stochastic game approach," in *Proceedings of the American Control Conference*, 2010, pp. 1059–1064.

[24] D. Gu, "A differential game approach to formation control," *IEEE Transactions on Control Systems Technology*, vol. 16, no. 1, pp. 85–93, 2008.

[25] Y.-C. Ho, "Team decision theory and information structures," *Proceedings of the IEEE*, vol. 68, no. 6, pp. 644–654, 1980.

[26] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. New York, NY: Springer-Birkhäuser, 2013.

[27] P. Trodden, D. Nicholson, and A. Richards, "Distributed model predictive control as a game with coupled constraints," in *Proceedings of European Control Conference*, 2009, pp. 2996–3001.

[28] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.

[29] A. Fabrikant, A. D. Jaggard, and M. Schapira, "On the structure of weakly acyclic games," in *Algorithmic Game Theory*. Springer, 2010, pp. 126–137.

[30] H. P. Young, *Strategic Learning and Its Limits*. New York, US: Oxford University Press Inc., 2004.

[31] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, 1994.

[32] D. Leslie and E. Collins, "Individual Q-learning in normal form games," *SIAM Journal on Control and Optimization*, vol. 44, pp. 495–514, 2005.

[33] D. Fudenberg and D. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[34] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. **14**, pp. 124–143, 1996.

[35] J. R. Marden, G. Arslan, and J. S. Shamma, "Cooperative control and potential games," *IEEE Transaction on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. **39**, no. 6, pp. 1393–1407, 2009.

[36] D. P. Foster and H. P. Young, "On the nonconvergence of fictitious play in coordination games," *Games and Economic Behavior*, vol. **25**, pp. 79–96, 1998.

[37] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff based dynamics for multi-player weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. **48**, no. 1, pp. 373–396, 2009.

[38] D. Leslie and E. Collins, "Convergent multiple-timescales reinforcement learning algorithms in normal form games," *Annals of Applied Probability*, vol. 13, pp. 1231–1251, 2003.

[39] S. Huck and R. Sarin, "Players with limited memory," *Contributions to Theoretical Economics*, vol. 4, no. 1, 2004.

[40] D. Leslie and E. Collins, "Generalised weakened fictitious play," *Games and Economic Behavior*, vol. 56, pp. 285–298, 2006.

[41] A. C. Chapman, D. S. Leslie, A. Rogers, and N. R. Jennings, "Convergent learning algorithms for unknown reward games," *SIAM Journal on Control and Optimization*, vol. 51, no. 4, pp. 3154–3180, 2013.

[42] E. Even-Dar and Y. Mansour, "Learning rates for Q-learning," *The Journal of Machine Learning Research*, vol. 5, pp. 1–25, 2004.

[43] G. Arslan, M. F. Demirkol, and Y. Song, "Equilibrium efficiency improvement in MIMO interference systems: A decentralized stream control approach," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 2984–2993, 2007.